Probability Theory

- Variables: The exact event being tested, but outcome not yet determined — Need to be Random.
  - X, Y, Z
- Sample Outcomes — Any individual potential outcome
  - O_i / ω_i
- Sample Spaces — The collection of all possible outcomes
  - $S = \{O_1, O_2, O_3, \ldots O_n\}$ / $\Omega = \{\omega_1, \omega_2, \ldots\}$
- Events — A collection of a subset of S
  - $A = \{O_1, O_3, O_5, \ldots O_n\}$ / $\{\omega_1, \omega_3, \omega_5, \ldots\}$
  - $A \subset S$ / Ω

A 1 2 3 4 5 6 7 8 9 10 J Q K

♡ ◇ ♤ ♧

drawing Q♡ is an individual outcome

drawing ◇ is an event

~~drawing any card is~~

all cards make up the Sample Space

$S = \{A♡, 1♡, 2♡, \ldots, K♡, A◇, \ldots, K◇, A♤, \ldots, A♧, \ldots\}$

$A = \{A◇, \ldots, K◇\}$

$\omega_{♧} = Q♡$

Probability of an outcome: $P(O_i) = P(\omega_i) = \frac{1}{|\Omega|} = \frac{1}{52} = P(Q♡) = 0.0192 = 1.9\%$

Probability of an event: $P(A) = \frac{|A|}{|\Omega|}$

$P(◇) = \frac{13}{52} = \frac{1}{4} = 25\% = 0.25$

We'll just use decimals

We can deal with 2 events, A and B

- Draw 2 cards with Replacement

	A♡	1♡	.....	Q♧	K♧
A♡	A♡ A♡	A♡ 1♡		A♡ Q♧	A♡ K♧
1♡	1♡ A♡	1♡ 1♡		1♡ Q♧	1♡ K♧
⋮					
Q♧	Q♧ A♡	Q♧ 1♡		Q♧ Q♧	Q♧ K♧
K♧	K♧ A♡	K♧ 1♡		K♧ Q♧	K♧ K♧

Size of Ω (Norm)

$|\Omega| = 52 \times 52 = 2704$

$P(Q♡ \text{ and } Q♡) = \frac{1}{2704} = 0.00037 = 0.037\%$

$P(◇ \text{ and } ♤) = \frac{13 \times 13}{2704} = \frac{169}{2704} = 0.0625 = 6.25\%$

$P(◇ \text{ and } ◇) = \frac{169}{2704} = 0.0625$

- Draw 2 cards without Replacement

$|\Omega| = 52 \times 51 = 2652$

$P(Q♡ \text{ and } Q♡) = 0$

$P(Q♡ \text{ and } A♡) = \frac{1}{2652} = 0.00038$

$P(◇ \text{ and } ♤) = \frac{13 \times 13}{2652} = \frac{169}{2652} = 0.0637$

$P(◇ \text{ and } ◇) = \frac{13 \times 12}{2652} = \frac{156}{2652} = 0.0588$

	340 text	Phillips text
Sample Space	S	Ω
Sample Outcome	O_i	ω_i
Event	A	A
Variable	X	X

?

Probability Rules

1. $0 \leq Pr(\omega_i) \leq 1$
2. $Pr(S) = 1$
3. Total probability of Disjoint (Independent) Events

$$Pr\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} Pr(A_i)$$

Independent Events

Conditional Probabilities: $P(A)$ when you already know B has happened

$$P(A \mid B)$$

If B has no effect on A, then A & B are independent.

$$P(A \mid B) = P(A)$$

$$P(B \mid A) = P(B)$$

$$P(A \cap B) = P(A)P(B)$$

- Rolling 2 dice
- Drawing 2 cards w/ replacement
- Flipping a coin

If B does have an effect on A, then A & B are dependent

- Drawing 2 cards w/out replacement
- Making Purchases & ~~having~~ income

Either way,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \qquad P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

Consider a study of people who smoke, relating it to lung cancer

		Cancer C=1	Cancer C=0	
Smoker	S=1	0.90	0.01	0.91
	S=0	0.03	0.06	0.09
		0.93	0.07	

$$P(S \cap C) = 0.90$$

$$P(S \mid C) = \frac{0.90}{0.93} = \frac{P(S \cap C)}{P(C)}$$

$$P(C \mid S) = \frac{0.90}{0.91} = \frac{P(S \cap C)}{P(S)}$$

~~Do Random Variables first (ch 3.4)~~

Continuous Variables

- Tables work for discrete variables, but not continuous
- We need calculus

Examples include
- Rainfall
- Height 68" → ~~[67.5", 68.5")~~ [67.5", 68.5")
- Time Class ends at 9:20 → [9:18, 9:22]

We can't do discrete probabilities as exact outcomes are unlikely

$$P(h = 68\text{" exactly}) \simeq 0$$

So, we use a density function, and we focus on probabilities that an outcome falls within a range.

The probability density function of a variable X is,

$$f_X(\omega_i) = \frac{1}{\sqrt{2\pi}} e^{-\omega_i^2/2}.$$

Not a probability. Known as a likelihood

pdf can be an number from $\emptyset$ through ∞

$$f_X : \Omega \to [0, \infty)$$

Graph in Python
conda install matplotlib
conda install scipy
or pip

$$P(X \in A) = \int_{\omega \in A} f_X(\omega)\, d\omega$$

cdf has to be a number between 0 & 1

$$P(X \in A) : \Omega \to [0,1]$$

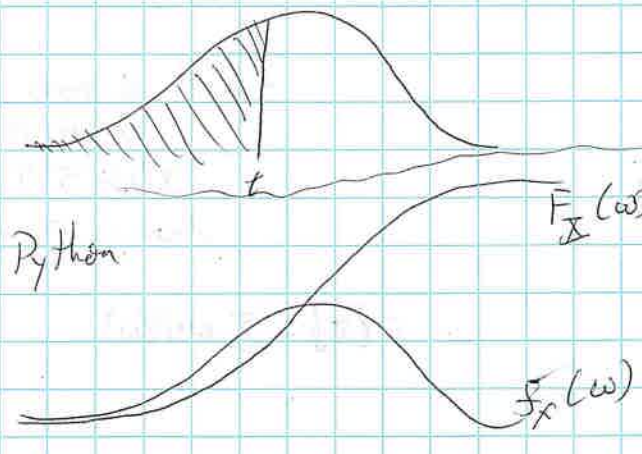
$$P(X \in [-1, 0]) = \int_{-1}^{0} f_X(\omega)\, d\omega$$

Area between boundaries of A.

Graph in Desmos

The Cumulative Density Function

$$F_X(t) = \int_{-\infty}^{t} f_X(\omega)\, d\omega$$



Graph in Python

Random Variables

A function that maps an outcome to an assigned value

Random Process → outcome → assigned value

A game:
- Roll a 1, Get 5 points
- Roll another odd, Get 1 point
- Roll an even, then lose 2 points

$\Lambda = \{1, 2, 3, 4, 5, 6\}$ Sample Space of possible outcomes

$\Omega = \{5, -2, 1, -2, +1, -2\} \in \mathbb{R}$ Sample Space of possible values

$X : \Lambda \to \Omega$
$X(\lambda) = \omega$
$X(4) = -2$
$X(1) = 5$
$X(3) = 1$

We often define variables s.t. Ω is a set of real values.

In Data Science, we rarely deal with a single value $\omega \in \mathbb{R}$ but a d-dimensional vector. Then,

$$\Omega = \{\omega \mid \omega \in \mathbb{R}^d\}$$

Do Retweet Discrete

Expected Values

$E[X]$ is the weighted average of Ω

- Biased coin? $P(H) = 0.4$ $P(T) = 0.6$ $\Lambda = \{H, T\}$
- Points: $H \to 5$ $T \to 1$ $\Omega = \{5, 1\}$
  $X(H) = 5$ $X(T) = 1$
- Roll 10 times. On avg., outcomes will be

$\lambda =$	H	H	H	H	T	T	T	T	T	T
$X(\omega) =$	5	5	5	5	1	1	1	1	1	1

Average $= \frac{20+6}{10} = \frac{26}{10} = 2.6$

Do in Python

$$E[X] = \sum \omega \cdot P(\omega)$$

$$E[X] = 5 \cdot P(H) + 1 \cdot P(T)$$
$$= 5 \cdot 0.4 + 1 \cdot 0.6$$
$$= 2.0 + 0.6 = 2.6$$

Expected values are linear operations:

$$E[X+Y] = E[X] + E[Y]$$

$$E[cX] = cE[X]$$

Example

Company ABC has 150 employees at an avg salary of $70k. Bonuses are given based on a 4-tier review:

	R=1	R=2	R=3	R=4
ω =	$1k	$2k	$3k	$4k
P(ω) =	0.05	0.30	0.40	0.25

(class example)

As a new employee, calculate your expected income

$E[S] = 70$

$E[Pay] = E[S + R] =$

$= E[S] + E[R]$

$= 70 + (1 \cdot 0.05 + 2 \cdot 0.30 + 3 \cdot 0.40 + 4 \cdot 0.25)$

$= 70 + (0.05 + 0.60 + 1.20 + 1)$

$= 70 + 2.85$

$= \$72.85k$

Variance

Written in terms of the expected value

The variance is a measure of how spread out the data is.

$$Var[X] = E[(X - E(X))^2]$$

$$= E[X^2] - E[X]^2$$

$E[X^2 - 2XE(X) + E[X]^2]$

$E[X^2] - 2E[X]^2 + E[X]^2$

$E[X^2] - E[X]^2$

As class Design Python programs for $E[X]$, $Var[X]$

$$\sigma_X = \sqrt{Var[X]}$$ → Same units as X

Covariance

How much 2 variables vary in relation to each other.

$$Cov[X, Y] = E[(X - E(X)) \cdot (Y - E[Y])]$$

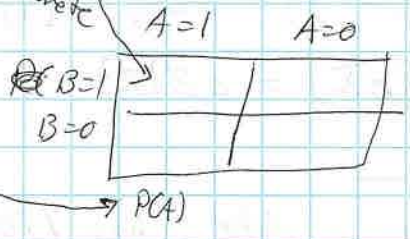
In lin alg:

$\bar{X} = \{x_1 - \bar{x}, x_2 - \bar{x}, \ldots\}$

$Cov[X, Y] = \frac{\bar{X} \cdot \bar{Y}}{n-1}$

Joint Probability: $P(A \cap B)$ → (Continuous) $f_{X,Y}(x,y) = P(X=x, Y=y) = P(x \cap y)$

Marginal Probability: $P(A)$



Conditional Probability: $P(A|B) = \frac{P(A \cap B)}{P(B)}$ → $f_{X|Y}(x,y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$

Discrete — Continuous

Bayes' Rule

You have some models and some data. We want to find the most likely model.
(M) (D)

$$P(M|D) = \frac{P(D|M) \cdot P(M)}{P(D)}$$ ← Sometimes

We can compute this

$$P(M|D) = \frac{P(M \cap D)}{P(D)}$$

$$P(D|M) = \frac{P(M \cap D)}{P(M)} \hookrightarrow P(M \cap D) = P(D|M)\,P(M)$$

You want to give an award to the best burger in town, based on
- Satisfaction
- Sales.

In one day, you find the following sales:
McDonalds: 120 burgers ← ? Great Sales,
Buster's Cafe: 75 burgers
Malt Shoppe: 55 burgers

Polling thier customers about satisfaction (come because it's good or because it's convenient):
McDonalds: 40% Satisfaction
Buster's Bistro: 70%
Malt Shoppe: 90% ← ? Great Satisfaction, but low sales.

If you see someone walking around with a burger and satisfied with it, what prob. is it that it came from ___ ?

$$P(M_c \mid S) = \frac{P(S \mid M_c) \cdot P(M_c)}{P(S)} = \frac{0.40 \cdot \frac{120}{250}}{P(S)} = \frac{0.192}{P(S)}$$

↑ Don't know. That's ok...

$$P(BB \mid S) = \frac{P(S \mid B) \cdot P(BB)}{P(S)} = \frac{0.70 \cdot \frac{25}{250}}{P(S)} = \frac{0.21}{P(S)}$$

$$P(MS \mid S) = \frac{P(S \mid MS) \cdot P(MS)}{P(S)} = \frac{0.90 \cdot \frac{55}{250}}{P(S)} = \frac{0.198}{P(S)}$$

In Data Science, we use this to find the model that most likely fits the data by finding how likely the data fits the model

Take a set of models Ω_M. We want $M \in \Omega_M$ that maximizes $P(M/D)$

$$M^* = \underset{M \in \Omega_M}{\operatorname{argmax}} \; P(M \mid D)$$

↑ Known as the Posterior (The result after the test)

What is argmax? The argument that maximizes the function.

$f(x) = \sin(x)$

$\max\{f(x)\} = 1$

$\underset{x \in [0,\pi]}{\operatorname{argmax}}\{f(x)\} = \pi/2$

likelihood ↓ Known as the Prior (some knowledge given prior to test)

$$M^* = \underset{M \in \Omega_M}{\operatorname{argmax}} \; P(M \mid D) = \underset{M \in \Omega_M}{\operatorname{argmax}} \; \frac{P(D \mid M) \cdot P(M)}{P(D)} = \underset{M \in \Omega_M}{\operatorname{argmax}} \; P(D \mid M) \cdot P(M)$$

$P(D)$ is known and fixed, so $P(D)$ is fixed and doesn't determine the argmax ←

If $P(M)$ is a constant, then $M^* = \underset{M \in \Omega_M}{\operatorname{argmax}} \; P(M \mid D) = \underset{M \in \Omega_M}{\operatorname{argmax}} \; P(D \mid M)$

Maximum Likelihood Estimator

~~3.8~~

$P(M)$ a constant? Is it?

In the example, we could have just looked at $P(S|BB)$, and it would have given a lot of information.
Multiply by $P(BB)$ just adds a little more information.
We do want to use $P(M)$, but if we don't know it, we can make a first estimate.

What exactly are M & D?

$D = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\} \in \mathbb{R}^n$ independent observations

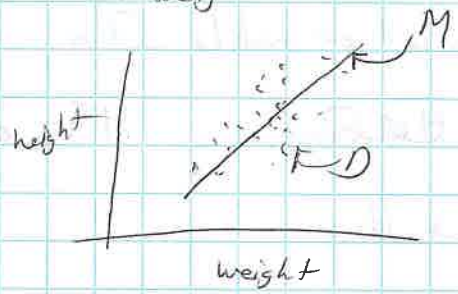
M: something that explains structure in the data

Example 1: I want to describe the heights of all students in class

$D = \{x_i\} \in \mathbb{R}^1$ (one value for each observation)
$M = \bar{x}$

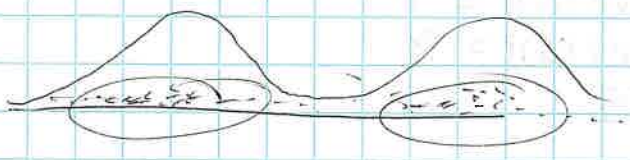
Example 2: I want to compare height with weight

$D = \{\vec{x}_i\} \in \mathbb{R}^2$
M: Linear Regression



Example 3: Clustering

$D = \{\vec{x}_i\} \in \mathbb{R}^d$
M: Set of points



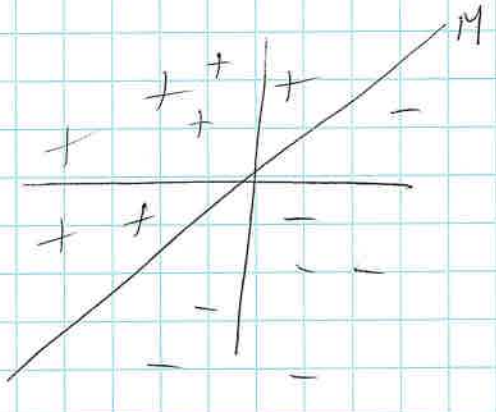
Example 4: PCA (Principal Component Analysis)

$D = \{\vec{x}_i\} \in \mathbb{R}^d$
$M \in F_k$ k-dimensional subspace

Dimensionality Reduction

Example 5: Classification (Linear)

$D \in \mathbb{R}^{d}$ $x \in \{-1, +1\}$



Let's do an example on Python

- Create an array of random numbers f/ normal dist. ($\mu = 0$, $\sigma = 1$)
- Make histogram
- Find $\bar{x}$ and s
- Create normal dist from data

$$f_x(\omega) = \frac{1}{s\sqrt{2\pi}} \exp\left(\frac{-(x-\bar{x})^2}{2s^2}\right)$$

In Python 3080_03.ipynb

- Plot hist & dist together
- Add normal dist to plot

pdf → like a probability, but not exactly. We can still use it

"P"$(x_i = x \mid M = m)$ ~ $P(D_x \mid M)$

We know the model. Given that, what is the probability that x occurs in this model?

$$P(D \mid M) = \prod_i \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$